Information Theory as a Measure of Sociodemographic Change

Zachary Lancaster, MCRP '18

Abstract

The current discourse around gentrification has been primarily focused on an economic understanding of gentrification. That is, that the gentrification can be primarily viewed through changes in housing prices in an area. That this increase is the sole and primary cause of displacement. While this economic explanation is an important component, it is not the only cause. The image, or perception of, a neighborhood changes as its socio-demographic and socio-economic circumstances change. As these circumstances change a process of replacement occurs and the social ecosystem begins to shift. This leads to the experiential phenomenon of gentrification, where the richness of social life declines as differing views and experiences are removed. This paper proposes a methodology for using metadata, specifically data concerning the entropy, or level of uncertainty or disorder, in demographic data to try to detect these changes and then applies this method to the City of Atlanta, in particular the areas around the Atlanta Beltline Eastside Trail.

Keywords: Statistics, metadata, gentrification, methodology

Introduction

Many of the City of Atlanta's neighborhoods are experiencing rapid change as downtown neighborhoods experience an influx of new residents seeking jobs in the booming entertainment and technology industries[1]. A testament to this is the growth in population of Fulton County, which includes the core of downtown areas of Atlanta and a majority of the city's administrative area and which has increased in population from 816,000 in 2000 to 996,000 in 2016[2] a growth of just over 18% in less than 20 years.

As with many cities experiencing re-densification of their downtown cores, Atlanta is experiencing gentrification, and there is a commonly held belief that this is a process of displacement due primarily to changes in housing costs. But gentrification is better defined as a

[1] *Atlanta City Design* (2017)
[2] United States 2000 Decennial Census; United States Census 2016 American Community Survey 5-Year Estimate

change in "social structures and housing markets"[3]. This phenomenon has caused a great deal of concern among citizens and denizens of the city alike. It is of particular concern for Atlanta given the City's troubled history with racial segregation in housing, services, education and other city functions[4]. Many in the city, such as the Housing Justice League and other activist groups, are concerned about the effects of displacement on local communities. There are growing concerns that neighborhoods such as Summer Hill and Peoplestown will be next on the long list of neighborhoods that are irrevocably changed due to development in the City of Atlanta.

The areas near Atlanta's 22-mile 'rails-to-trails' project, The Beltline, has been a lighting rod for controversy[5]. Research conducted by Immergluck has shown clear evidence that, as the project has advanced through the city, it has created significant negative externalities and placed them on the existing population near The Beltline[6]. As an example, housing prices have exploded in recent years with home prices in the Old Fourth Ward, a historically African American neighborhood adjacent to The Beltline's oldest segment The Eastside Trail, rising significantly in the years between 2010 and 2015. Many in the city are concerned, as the project continues to advance through the city, that this pattern will repeat, leading to disrupted communities. South Atlanta near the Bell Quarry, and neighborhoods in the East Side and West End are already starting to feel the early effects of speculation related to the encroachment of The Beltline. The rupture in the continuity of a community caused by rapid displacement of an extant population is being shown to have negative public health outcomes, an emerging fact that cannot be overlooked.

Up to this point we have defined the phenomenon of gentrification almost exclusively as an economic outcome of rising housing costs. However it is not just price that drives gentrification; there are significant and complex social and phenomenological factors[7] involved as well.

---

[3] Ruth Glass defined gentrification in 1964 in *London: Aspects of Change* as "the [changes] in social structures and housing markets [in an area]".
[4] D. Sjoquist, *The Atlanta Paradox* (2000) p. 88-127 covers in good detail the problems of residential segregation in Atlanta.
[5] B. Powers "Putting the Breaks on Runaway Gentrification in Atlanta" *Citylab* Nov 11, 2017; https://www.citylab.com/equity/2017/11/putting-the-brakes-on-runaway-gentrification-in-atlanta/545555/ accessed 07/31/2018
[6] D Immergluck, T Baylan. "Sustainable for Whom" *Urban Geography* (2017) vol 39, p.546-562
[7] C. Allen. "Gentrification 'Research' and the Academic Nobility: A Different Class?", *International Journal of Urban and Regional Research* 2008 (vol 32, issue 1)

Anecdotally, gentrification tends to be perceived as an observed loss of diversity and not a problem purely of economics. While it is possible to describe the phenomenon of gentrification in economic terms, such as the explosion of housing prices and the displacement of a local population due to this factor, metatextual analysis of comments made by planning students, professionals and faculty tends to indicate that these are indicators of a phenomenon, ones that could be measured and represented, but not in-and-of themselves that cause. Rather gentrification is a phenomenon, a process, where people perceive in an area a decline in the types of (other) people they encounter and in turn the experiential qualities of a neighborhood or place[8]. What this implies is that there is a deeply complex process that goes into determining how the consensus of the zeitgeist chooses to organize an area. Is it gentrified or ghettoized, is it a business district, entertainment or arts? This goes beyond the dollars and cents realities of property values and is less as functions of how authorities wield zoning to shapes an area as they do from the public narrative that forms from the experiences and interactions available inside of that area[9]. While this initial work deals only with the phenomenon of gentrification it is easy to envision extending similar ideas to other classifiers. The sociological phenomenon of gentrification is in need of investigation. This paper intends to begin to describe and represent this in quantitative terms, recognizing the short comings of this approach.

What is clear is that gentrification cannot be described as a one or two variable problem; it is a multidimensional, highly correlated socio-demographic and socio-economic phenomenon. It is more likely the case that it is better described as a complex set of socio-demographic and socio-economic variables interacting across time and within space to alter, in the words of Kevin Lynch, the "image" of that place[10]. In her own attempts to describe the processes at work in Greenwich Village in New York City in her seminal work *The Death and Life of Great American Cities*, Jane Jacobs called the city, its growth and the way it changed over time, a problem of

[8] This observation is derived from conversations with students and professionals in the planning profession asking them to define and describe the process of gentrification between November 2017 to April 2018. These were taken as informal interviews, and do not represent an exhaustive study. A more rigorous, and less biased, study is required to better define the phenomenon, but that work laid outside of the scope of this paper.

[9] A tongue and cheek article *The Four Horsemen of Gentrification* by Zain Khalid published in 2015 provides an example of how this is as much a social phenomenon as it is an outcome of some purely economic externality. https://www.mcsweeneys.net/articles/the-four-horsemen-of-gentrification

[10] K Lynch, *Image of the City* described the image of the city we suppose here that this extends beyond the physical into the ephemeral qualities of interpersonal social interactions

"organized complexity"[11]. She observed that cities are complex systems in which the factors that govern their growth are not all, or even necessarily all measurable, understood but could be understood through analysis of a small number of attributes (variables) that dominate the process, a view that was inspired by work in complexity theory at the time, stating that "variables are many, but they are not helter-skelter; they are 'interrelated into an organic whole'."[12]. A city is, as Jacobs indicates, the collective, independent yet interrelated actions of a vast number of agents, individual and institutional, that ultimately give rise to an emergent physical, economic and sociological state. These systems have a large number of variables, but, much like other natural systems, have only a few dominant attributes that influence the greater system behavior. One of these factors is the diversity, or richness, found in the city's population as a whole, both locally and across local boundaries.

The process of gentrification has been ethnographically explored, and people have sought to document the experience of displacement[13]. There has also been significant statistical work to describe the absolute outcomes of gentrification as a force of the housing market[14]; what is missing is a quantitative description of the how gentrification begins to impact the "Image of the [Social] City"[15]. In the spirit of Emelie Durkheim, we intend to approach this topic with scientific rigor[16] and begin to develop a method for describing this image by way of understanding socio-demographic metadata. First, we intend to set a foundation for understanding social systems as analogous to natural ecological systems, and a foundation for the specific analytical method that will be applied, that of entropy analysis. We will then describe both the study area and its importance within the discourse of gentrification within Atlanta and will present in detail the analytical method that we are seeking to develop. Finally, we will discuss the interpretation of the results, further development and its place within the discourse around the phenomenology of gentrification.

Background and State of Knowledge

[11] J Jacobs, *Death and Life of Great American Cities* p. 429
[12] J Jacobs, *Death and Life of Great American Cities* p. 430
[13] M Desmond *Evicted* (2016)
[14] D Immergluck, T Baylan. "Sustainable for Whom" *Urban Geography* (2017) vol 39, p.546-562
[15] K Lynch, *Image of the City* (1960)
[16] E Durkheim, *The Rules of Sociological Method* (1895)

To begin we require a lens through which to view and make sense of the complexity of a city as a complex social system. We propose that this lens can be developed through the study of ecology as it provides an appropriate analogy and allows us to derive a set of metaphors that appropriately frames and contextualizes the problem. Ecology, here, deals with the study of natural systems and the interactions among organisms and their environments. While the history of ecology stretches back to the Greeks, it was not till the 19th century that ecology became a field of study in its own right, and not until the mid-20th century, through the likes of landscape architect Ian McHarg, that it would make its way into the discourse of cities[17]. Today urban ecology is a growing field of study that tries to describe how inhabitants of a city interact with their environments -- physical, social, political and infrastructural[18]. In this context we seek to use ecology more in its classic sense, as a physical science, to frame the problems of gentrification.

Systems of plants, such as those found in wetlands, are part of highly diverse biological systems in which plants and animals are dependent on each other for survival in a delicate symbiosis. In these systems, native species of flowers, grasses, plants and animals flourish in concert with one another, each species nested within in a complex web of interrelationships that are difficult to understand in their totality. These systems tend to be stable within their given epochal and secessional regimes, shifting and changing in relatively predictable patterns. Shifts in key system health indicators, even subtle shifts, can have large negative impacts on the disparate threads within the ecosystem. Introductions of new species, or changes in nutrient availability or changes in the external environment can cause reductions in species diversity (or richness) within the system and potentially, if left unaddressed by some mechanism, lead to system fragility and eventual collapse[19]. These systems benefit from a diversity of species being present in the system, and in ways in which there is both redundancy and resiliency in the system. This is consistent, by way of metaphor, with the definition provided by ecology, and therefore opens up the possibility of applying methods from ecology to better understand the processes of gentrification through the lens of this metaphor.

[17] I. McHarg *Design With Nature* (1968)
[18] C Waldheim *Landscape as Urbanism* (2016)
[19] A Maguran *Ecological Diversity and its Measurements* (1988)

The study of diversity, within ecology, is measured quantitatively through diversity indices. There are many and a full review of these methods can be found in many ecology texts[20]. In these methods, the goal is to represent in some way the richness of speciation within a community. While there are differing approaches with interpretive differences, due to the mathematical properties of each approach, they are each similar and provide similar information concerning the richness and diversity of an ecological system. This paper will primarily focus on the Shannon diversity index[21], a derivation of information entropy methods Shannon proposed in the early 20th century in his *Mathematical Theory of Communication*[22]. This method provides a general and easily understandable representation of the richness and complexity of an ecosystem.

This method makes use of two useful values. The first is entropy (disorder or uncertainty) which equates to a measure of abundance within the system. Put simply, it abstractly represents how unsure one would be of the type of species they would find if taking a random sample at a random time. The other useful value is equitability (evenness of distribution) which shows how well mixed these populations are. This provides a good view of the overall health of the system from a high level. When looked at over time, it provides a means for understanding both the presence and distribution of species within the system. While this method is by no means exact, it is a good measure by which to measure the overall health of a system as a function of the diversity of species present within its boundaries.

[20] A Maguran *Ecological Diversity and its Measurements* (1988) p 7-45 serves as a good introduction to methods of measuring species diversity within the field of Ecology.
[21] This index is described and criticized in the context of Ecology in L. Joust, "Entropy and Diversity", *Oikos* (2006)
[22] C. Shannon, *A Mathematical Theory of Communication*

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

p_i: probability of a specific value within a signal

Equation 1: Shannon's Entropy Index

$$E = H/H_{max} = H/\ln S$$

S: number of species present in the population

Equation 2: Shannon's Equitability Index

We can, by way of analogy, view social systems in a similar way, as a kind of social ecology[23] where a "species" is one or a collection of demographic traits that represent the richness of the social experiences found in an area. By configuring our research in this way, we are able to take advantage of the tools and techniques of ecology to study our human communities, and the richness of diversity within. This use of ecology as an analogous theoretical framework is not novel. In the early 20th century McKenzie, a part of the Chicago School, took notice of these same analogies in his 1924 paper *The Ecological Approach to the Study of The Human Community*. McKenzie proposed a human ecology, proposing that society can be understood by analyzing, systematically, the spatial and temporal relations of humans.[24] This approach was developed because, as Catton and Dunlap (students of McKenzie) argue, sociology had to that point had "focused on humans to the neglect of habitat…"[25] and that sociology had left out the "object world". Critics of human ecology point out that this method inverts the relationship, focusing on a biological determinant and missing the psycho-social aspects of culture and human communities. Jean Bernard stated that human ecology excludes "culture, attitudes, and habit patterns; personality and social types…community conscious; and … such phenomena such as

[23] Not to be confused here with the Social Ecology studied by the Vienna School of Social Ecology whose primary concerns tend to be anthropological, historical and economic studies of the rise of societies writ large. Here we intend the term to be a more metaphorical concept, one that encapsulates the similarities between human social communities and plant/animal communities.

[24] J T McKenzie "The Ecological Approach to the Study of The Human Community" *The Journal of Sociology* (1924) 30 no 3 p. 287-301

[25] W Catton, R Dunlap "Environmental Sociology: A New Paradigm" *American Sociologist* (1978) 13 (1): p41-49

interaction both interpersonal and intergroup." in its pursuit of its biological description of human communities.[26]

We agree that human ecology is insufficient for the development of a metanarrative about human systems. As a theoretical framework, it suffers from relying too much on the analogy and does so at the expense of the differences between humans and non-human communities, such as psycho-social influences, the fluidity of human movement, and the human ability to reshape its environment. Human ecology suffers from placing the physical built environment above and before any cognitive model of that same environment. Human ecology fails to account for the psychogeography a city, instead relying on a model based on human geography. We would like to propose that, while the physical environment and the qualities of that environment are important, these changes to the physical environment are, like the economic explanations of human behavior, outcomes of change in the psychogeography of the city. Changes in the social environment impact the psycho-social landscape, changing the perception of a part of the city. It is not possible to explain the larger phenomenon of how humans and their interaction shape how people conceive of a place purely through physical affect. What human ecology does provide is a convincing argument that there are enough similarities between human communities and non-human plant and animal communities for us to operate the tools of ecology analogously so long as it is recognized that, as an analogy, these tools do not produce an explanatory model, but rather an abstraction – a representation – that can be used for communication about a phenomenon.

With this limitation in mind it is noteworthy that the social sciences have taken advantage of the tools of ecology for some time to create these kinds of representations. Tools such as isolation, concentration and diversity, used regularly in the study of ecology, are deployed in support of the study of human systems. A good meta-analysis of common classical methods was performed by Massey and Denton in 1988 in their paper *The Dimensions of Residential Segregation*. Their paper serves as a good overview of the general properties and applications of these kinds of

[26] J Bernard "Social-Psychological Aspects of Community Study: Some Areas Comparatively Neglected by American Sociologists" *British Journal of Sociology* (1951) 2 (1) p.12-30

indexes and methods in the social sciences[27]. The issue with the application of these methods so far has been largely one of limiting scope in terms of application and interpretation. These tend to be univariate models, ones that fail to capture the complex and emergent phenomena that define a city. The original models though are multidimensional, the Boltzmann entropy model requiring 6 dimensions, and the limiting of these indexes to single dimensions seems unwarranted mathematically.

Applications of Entropy as a Measure of Diversity in the Social Sciences

The application of statistical entropy methods in the social sciences has met with success in investigation of specific phenomena, such as segregation in school systems in the United States of America or the diversity of land uses in relation to transit alternatives. Entropy analysis has served to demonstrate and communicate changes along single variables. We would like to bracket the work done on the use of entropy within the social sciences by looking at two specific examples: Theil and Finizza from 1972 and by Barner et al. in 2017. We choose to look at these in specific as they bracket key points in the discourse of entropy in the social sciences. Theil and Finizza represent an early implementation, with Barner et al demonstrating a more sophisticated and contemporary methodology.

In the case of Theil and Finizza they were investigating segregation in school districts in the 1970's in the wake of the civil rights era. This method was first proposed in Theil's 1972 text *Statistical Decomposition Analysis* and expanded in a paper published later in the same year[28]. This paper demonstrated how entropy can be used to measure the level of segregation within a school district. This was being used to try to measure to what extent segregation and integration were occurring at the time. Theil's method is sound, for its intention, measuring, in a base two logarithm, proportions of the majority group to its proportional compliment. This created a concave, symmetric outcome that increased to a point where the majority (in this case white) students comprised an even portion of the population. Theil and Finizza's method is also aspatial and criticized by Massey and Denton for suffering from invariance, an artifact of its binary

[27] D. Massey, N. Denton, "The Dimensions of Racial Segregation", *Social Forces* (1988) 67 (2)
[28] P. Theil *Statistical Decomposition Analysis* (1972); P. Theil "A Note on the Measurement of Racial Integration of Schools by Means of Informational Concepts" *The Journal of Mathematical Sociology* 1 (2) p.187-1983

nature. Theil and Finizza showed that it was possible to create a statistical aspatial comparison of how well the districts were performing both individually and in relation to one another. This representation is an easy and effective indicator to represent proportionality and general mixing within school districts. While Theil and Finizza did not dive deeper into the underlying causes, instead choosing to focus on demonstrating where segregation was occurring, it did demonstrate that it is possible to measure population segregation, or a strongly non-uniform distribution, as possible in socio-demographic problems.

Barner et al, on the other hand, represents a more contemporary method and an attempt to get at deeper questions. Barner was able to take advantage of a new set of computational tools to allow for a larger scale analysis of urban structure that focused on the emergence of the physical form of the city. The goal of Barner's paper, *Multiscale Entropy in the Spatial Context of Cities,* is to describe how, spatially, London has reached its current state through the aggregation of small bottom up changes by attempting to maximize the entropic change across time and space. This method has been discussed for some time in literature around urban simulation[29] and is demonstrated by Barner, the University College of London Urban Center for Advanced Spatial Analysis, and his collaborators from other institutes. Their model demonstrates how the process of seeking entropy maximization, the maximizing of a metaphorical free energy, gives rise to the patterns of urban development. This created a multiscale spatial model that is consistent with "both entropy in its statistical mechanics, but also an understanding of cities as complex systems.".[30] The method presented in their paper is robust and relates changes at differing scales as a way to describe larger emergent patterns of change induced by smaller complex coupled systems. It is, as Barner puts it, a measure that "relates the fast, microscopic behavior of the elements of a system to its slow macroscopic state."[31] Much of the work in contemporary applications of entropy maximization and entropy theory, Barner and Batty included, have to do with economic decisions and how those decisions impact the formation of urban patterns.[32]

[29] A derivation of these simulation methods can be found in M Batty *The New Science of Cities* (2013) p271-281
[30] Barner et al "Multiscale Entropy in the Spatial Context of Cities" *Physics and Society* (2017)
[31] ibid
[32] Y Zhang "Analysis of Urban Ecosystems Based on Information Entropy" *Ecological Modeling* (2006) 197 p1-12

This economic approach suffers from the same problem McKenzie and his human ecology framework did. It preferences a single modality that ignores the psycho-social aspects of the agents that make decisions within the urban system. While in the case of Barner investigation it may be argued that these are institutions making decisions at the scale of the parcel or block, it is also true that those decisions are not made in a sociological vacuum. What both works do is demonstrate how these tools can be interpreted in pursuit of an indicator. Theil and Finizza demonstrate how we can take advantage of the statistics of ecology to obtain a high-level understanding, while Barner, Batty and Zhang (taken together) represent attempts to show how entropy can be used to understand processes of change.

What we hope to do is to learn from the whole breadth of this nearly century long exploration by using well tested methods and applying them in a way that leads towards a mixed method that can explore complex societal narrative phenomena related to gentrification. We, for the time, leave out complex multiscale dynamics and the interplay of spatial regions and instead focus on key variables that describe in very general terms the demographic (race, age, and citizenship status) and economic (income, education, and household structure) richness of an area; items of concern when discussing gentrification.

Methodology

As has been demonstrated, there is a history in the last 50 years of applying methods of statistical entropy estimation to the study of urban problems and urban complexity. While there are an array of options in the social sciences,[33] we will make use of the most fundamental method with an intention of working outward in further work towards more complex methods. The key problem in this paper, as stated, has to do with relating entropy not across space or scale (which have been demonstrated by Theil and Finizza and Barner respectively), but to create a highly multidimensional measure that describes – quantitatively – gains or losses in the complexity (level of heterogeneity) of social environments across time. These changes lead to losses or gains

[33] D Massey, N Denton. 'The Dimensions of Residential Segregation', *Social Forces* (1988) 67 (2)

in social contact that can have differing impacts on socio-economic outcomes[34], but ultimately change the experiences available to people and thus reshape the image of that portion of the city.

Our method of choice to begin to describe this phenomenon is a Shannon entropy index, as an empirical measure of social complexity measured as a measure of disorder or uncertainty. This index is turned into metadata, an index about demographic datasets concerning the information content of a given variable. This index represents level of disorder (or randomness) within a closed (geographically bounded) system as a ratio of the maximum amount of possible uncertainty, which is achieved when the variable is uniformly distributed between levels. This method provides specific interpretive and mathematical advantages while keeping methodological complexity to a minimum. The results of the analysis will be used to form a descriptive study of the ways in which changes in patterns of social and economic demographic indicators describe changes in the complexity of social systems in the city limits of Atlanta that lay within Fulton County between the years 2010 and 2016, the period since Atlanta's Beltline project has been under construction[35], with special attention paid to areas surrounding The Atlanta Beltline Eastside Trail.

Study Area and Data Set

This study collected data at the census tract level for areas within the municipal boundary of the City of Atlanta within Fulton county. The data used as a part of this analysis are products retrieved from the United States of America Census Bureau. Samples were taken from both the United States of America 2010 decennial census[36] and the 2011 to 2016 American Communities Survey five-year estimates (ACS5)[37]. Variables for this were selected to cover two broad categorical identifiers, social identifiers and economic identifiers, with the belief that these varied independently and that there would be marked differences in how the ecosystem responds. Social identifiers attempt to describe, broadly, the ways that people create social groups, and

[34] M Hewstone "Neighborhood Ethnic Diversity and Trust: The Role of Intergroup Contact and Perceived Threat" *Psychological Science* (2014) 25 [3] p 665-674
[35] https://beltline.org/progress/progress/project-history/
[36] United States of America Census Bureau 2010 decennial census
[37] United States of America Census Bureau American Communities Survey

include identifiers such as age, race, and citizenship status. Economic identifiers are variables that describe the ways in which people organize themselves by class, household structure, education and income. Many of these variables are only available in tabular form at the census tract level, limiting in this study the spatial resolution available.

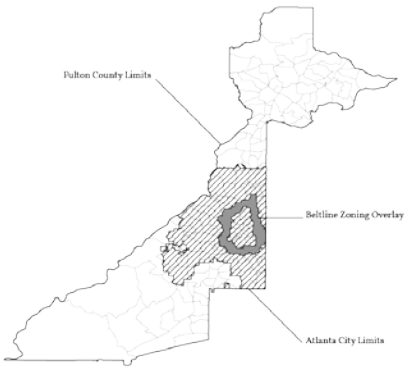

Figure 1: Fulton County Georiga with highlighted areas of interest

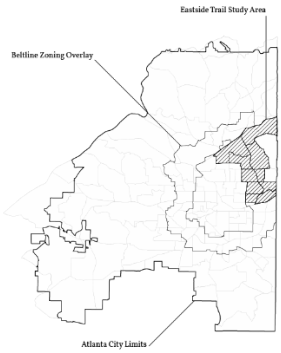

Figure 2: Atlanta City limits with highlighted study area

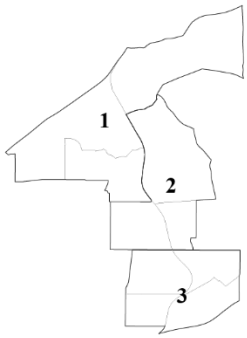

Figure 3: Study area specific areas represent (1) Ansley Park (2) Virginia Highlands and (3) Old Fourth Ward

The study area has shown significant growth over the study period. This growth can be attributed to rapid technical[38] and entertainment related job growth[39] and a general nationwide trend towards re-urbanization among millennials and retiring baby boomers[40]. These factors have led to growth within the city limits of nearly 50 thousand in the six-year study period.[41] Growth has not been evenly distributed across the whole of the city limits, as seen in figure 2. The Atlanta Beltline project has attracted significant development. The Eastside Trail, the oldest of the segments, which broke ground in 2010 and completed the first three and a half miles in 2011 has seen more than 10 major development or adaptive reuse projects completed or break ground

[38] Atlanta Business Chronicle "Growing our Region as High Tech Hub" https://www.bizjournals.com/atlanta/print-edition/2011/10/28/growing-our-region-as-high-tech-hub.html?page=all accessed 2018/06/30
[39] Atlanta Business Chronicle "Georgia's Film and Television Industry Economic Impact Reaches $6 Billion in Fiscal 2015" https://www.bizjournals.com/atlanta/news/2015/07/09/georgia-s-film-and-television-industry-economic.html access 2018/06/30
[40] The Atlantic "More Americans Moving to Cities, Reversing the Suburban Exodus" https://www.theatlantic.com/national/archive/2014/03/more-americans-moving-to-cities-reversing-the-suburban-exodus/359714/ accessed 2018/07/01
[41] United states census ACS5 estimate

during the study period. Analysis of the data set has therefore paid special attention to this area to look into how the richness of demographic information has changed during the course of the development and rapid subsequent gentrification of the area.

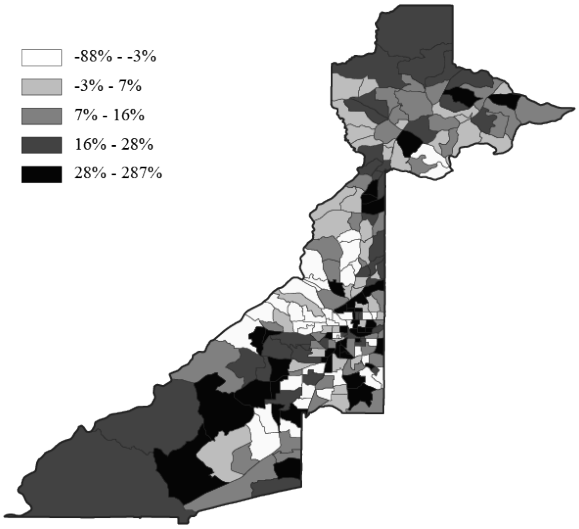

Figure 4: Population Change in Fulton County

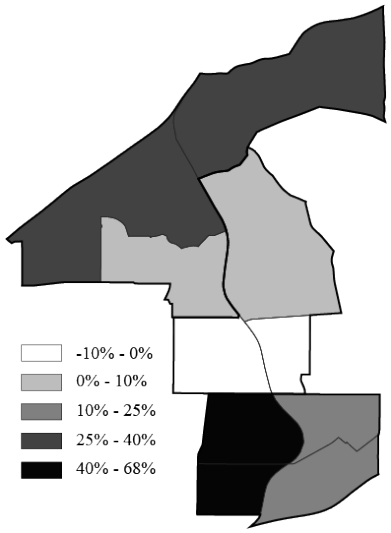

Figure 5: Study area Population

Variables were chosen based on a review of the United States Center for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI)[42]. The SVI is used to represent to what extent a community is able to prepare for and recover from disaster.[43] Generally speaking, marginalized communities are most at risk, and areas where those communities are concentrated would therefore rank higher. This index represents a standard measure used by many agencies who track at risk communities, such as immigrants and those living in poverty and their readiness and resiliency. This provides a metric to which we could compare, since many areas with high social vulnerability would contain a high proportion of people who would also be at risk for displacement during the process of gentrification.[44,45], since we assume, given the

[42] CDC Social Vulnerability Index https://svi.cdc.gov accessed 2018/05/02
[43] CDC Agency for Toxic Substances and Disease Registry "SVI Fact Sheet"
[44] H Pearsall. "From Brown to Green? Assessing Social Vulnerability to Environmental Gentrification in New York City" *Environment and Planning C: Politics and Space* (2010)
[45] The dataset may be obtained through ResearchGate doi:10.13140/RG.2.2.35617.92003 and 10.13140/RG.2.2.24712.72966

parallels to ecology, that ultimately, a highly resilient urban system is one that is comprised of the largest number of possible communities.

This study breaks the variables, found in table 1, into two broad classifications: social identifiers and economic identifiers. These, as mentioned, represent two different axes that we believe to vary independent of one another. The social variables, age, race, and citizenship status aim to capture the ways in which people form social groups based on a demographic basis. The economic variables, per capita income, educational attainment, and household structure seek to represent the ways in which class plays a role in the formation of social groups.

	Variable Name	Description and Census Identifier
Demographics	Age	Age broken out by sex (B01001)
	Race	Top coded race categories (B02001)
	Nativity & Citizenship Status	Place of birth by nativity and citizenship status (B05002)
Economics	Household Income	Household income in 2010 adjusted dollars (B19001)
	Household Size	Size of both family and non-family households. (B011016)
	Educational Attainment	Level of educational attainment for population (B15003)

Table 1: Variable description and provenance

Defining the Entropy Index

While there are multiple means of estimating an entropy statistic within physics[46] and for measuring species richness in ecology (e.g. The Simpson index), for our purposes we will focus on the Shannon entropy index. This index represents the most fundamental formulation of an entropy estimator, making the statistic both the most straight forward to calculate, and most

[46] See R Liboff. 'Gibbs vs Shannon Entropy' *Journal of Statistical Physics* (1974) 11 issue 4 p.343-357 for a review and comparison between these methods as applied in statistical physics.

easily modified to support higher dimensionality[47]. The value (E) is both easy to interpret and is represented in a form that can be used to represent change over time. This gives an index with advantageous properties for developing a metadata metric to describe changes in social systems.

It is important, though, to discuss the properties and specific ways in which we define and interpret the index that may be different from classical methods. The Shannon entropy index has properties that provide both statistical and interpretive benefits. The entropy value (*H*) is parabolic (see figure 1), meaning that it cannot be read for directionality, only the absolute distance from a maximum (1), or minimum (0) is able to be measured. This fact is carried over into the final value of the final equitability index (*E*) which is a ratio between the estimated entropy and the maximum possible entropy value (as described in equation 2). This index value can be interpreted as a distance from an ideal uniform distribution between all levels of a variable[48]. This gives the equitability value consistency across variables, regardless of the number of levels, making aggregation across variables possible without scaling and allowing the statistic to remain unbiased. These distances are averaged together giving a general idea of how, across all variables in a given dataset, evenness, or a measure of heterogeneity, of the information content. In short, the more heterogenous (closer to 1) the index is, the more information is contained. This means that the level of uncertainty is increasing, a person in that area is less and less likely to know whom they may encounter. This is broadly interpreted as a 'rich' social environment with a diverse range of potentials for social interactions with people within the environment who share different experiences and views.

The index is computed in multiple steps. The first step, equation 3, computes a value for *H* across a single variable accounting for all levels, *S*. It should be noted, as is displayed in theorem 1, that if the value of p_i, a percentage of the population in level or category $X_i | X_i \in X$, are approximately evenly distributed, which is to say $p_i \sim 1/S$, then the sum (*H*) will equal the natural logarithm of *S* (*ln(S)*).

[47] The original Boltzmann method, introduced in 1872, supposed a 6-dimenional phase space for an individual particle.

[48] The ideal distribution defined as the case where the cardinality of all discrete finite subsets within the set are equal: formally stated as $|X_i| = |X_j| \forall j \neq i$.

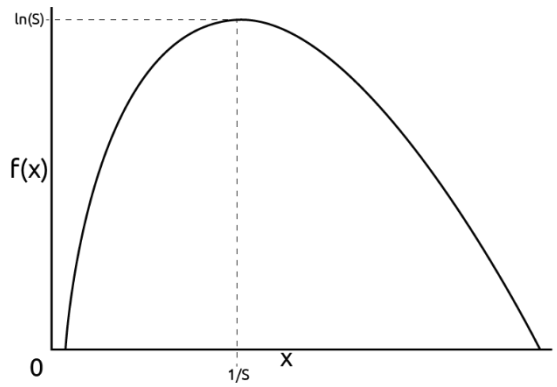

Figure 3: Curve for equation 3

$$H = -\sum_{1}^{S} p_i \ln p_i \; \forall \; p_i > 0$$

Equation 3: Shannon Entropy Index

$$\lim_{p_{i \to 1/S}} H \to \ln S$$

Theorem 1: The limit of the entropy function is the log of the cardinality of the set of subsets (number of levels)

The second step computes variable specific Shannon equitability (E). The equitability value is the ratio between a specific value *H* and the maximum value of *H* (H_{max}), shown in equation 4. It is simple to show that $H_{max} = \ln S$, which is consistent with theorem 1, and thus the maximum value of E is 1. This gives E a range of 0 to 1 by virtue of comparison to this maximized value given that H <= H_{max}. This leads to theorem 2, that as H that approaches its limit (lnS), as demonstrated in theorem 1, E monotonically approaches a limit of 1.

$$E = \frac{-\sum_{1}^{S} p_i \ln p_i}{H_{max}}$$

$$E = \frac{H}{\ln S}$$

Equation 4: Shannon Equitability Index

$$\lim_{H \to \ln S} E \to \frac{\ln S}{\ln S} = 1$$

Theorem 2: As H approaches the maximum value (see theorem 1) E approaches 1

Finally, these variable-specific values are averaged together to create the final indicators E_{demo} and E_{econ}. The final algorithm used to compute these values can be found in equation 5. As an average this function inherits the range of E, 0 to 1. It is possible to discuss these values in

relative terms both across system boundaries and longitudinally across time. Generally when comparing systems, values such that $E_{var} \sim 1$ are considered to indicate a system with more richness than another there $E_b < E_a <= 1$ would indicate that E_a has more richness, or demographic uncertainty, than E_b. In longitudinal or time series analysis positive values would indicate a system with that is expanding socio-demographic and/or socio-economic diversity. While a negative value would indicate a system that is currently in "decline" or contraction, where diversity and richness are declining.

$$E_{var} = \frac{1}{n}\sum_{j=1}^{n}\frac{-\sum_{i=1}^{S_j} p_i \ln p_i}{\ln S_j}$$

Equation 5: Shannon Equitability value for a given indicator

Analysis

As discussed earlier, the purpose of this index is to provide a method to represent metadata about these demographic data sets, that is data about data, to describe the patterns that lead to a perception of gentrification. What we know concerning the study area (shown in figure 2), is that after 2011, when the first sections of The Beltline were completed, the area started to experience rapid change. Major development and redevelopment projects began to reshape the physical environment, but more importantly the new amenity began to attract people who wanted to take advantage of it well before any of the major developments were complete. This new interest has led to major notable transformative projects in the study area including Ponce City Market, Old Fourth Ward Park (and its attached AMLI apartment developments). The tract containing Old Fourth Ward Park saw growth of nearly 70% population growth over the course of 2010 to 2016 growing from 2553 to 4278 residents[49] (this can be seen in figure 5).

[49] Tract GeoID 13121001700

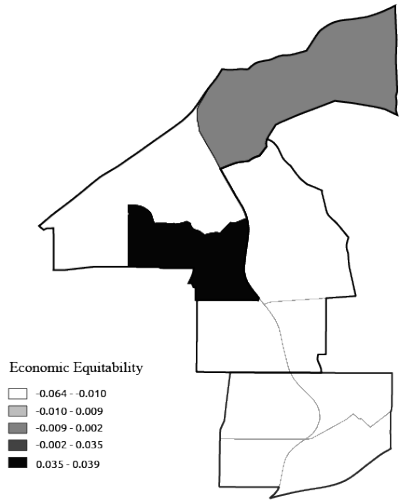

Figure 5: Ansley Park Economic Equitability

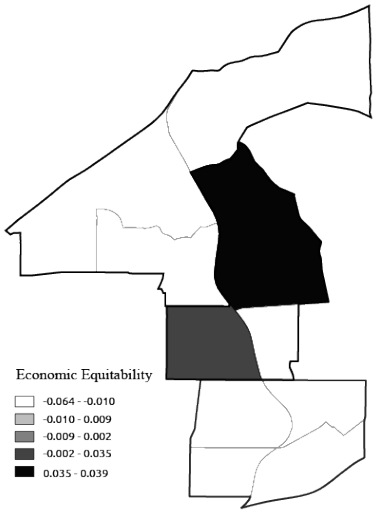

Figure 6: Virginia Highland Economic Equitability

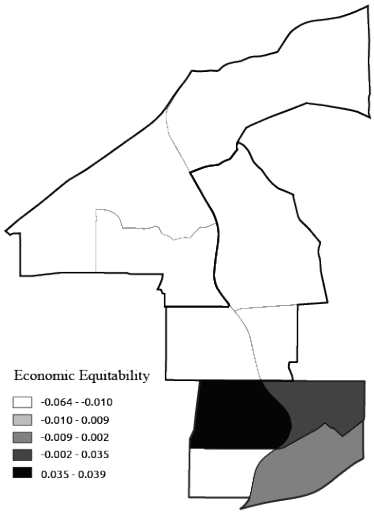

Figure 7: Old Fourth Ward Economic Equitability

Tract ID	Economic	Demographic
13121000500		
13121000200		
13121001400		
13121002900		
13121009200		
13121001700		
13121003000		
13121000400		
13121001300		
13121001600		

Table 2: Visualization of Economic and Demographic Slopes

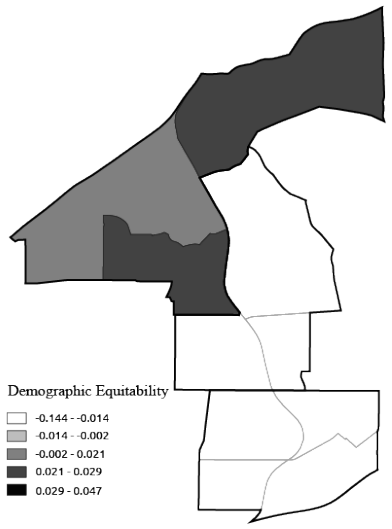

Figure 8: Ansley Park Economic Equitability

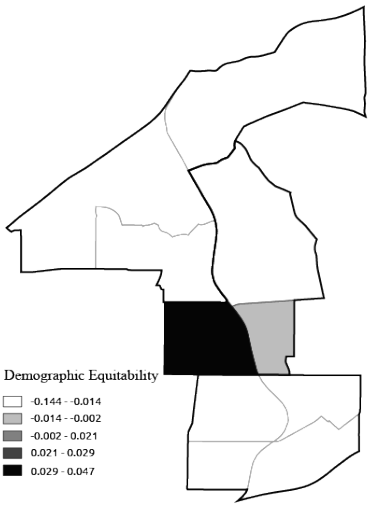

Figure 9: Virginia Highland Economic Equitability

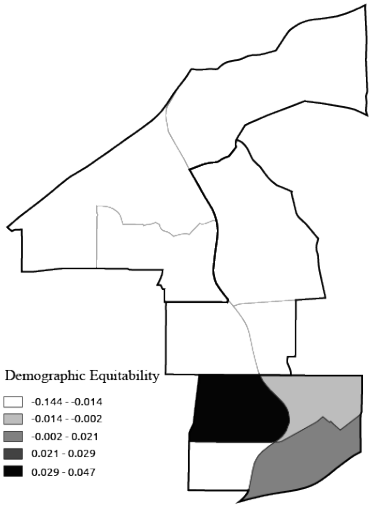

Figure 10: Old Fourth Ward Economic Equitability

What is inferred in figures 5 through 10 is that this process of growth has, when taken at a high level and at a coarse grain, caused an overall reduction in the Shannon Equitability measure (E). This finding follows the conventional wisdom surrounding gentrification and The Beltline. While there are areas where there has been growth, there are more areas that have seen declines in the richness of demographic and economic environment.

This reduction exhibits an interesting pattern when taken over the whole of the area. This is seen in the parabolic shape of the measures when graphed over time, figures 11 and 12. If we consider for a moment a theoretical expectation this is the expected outcome, we would expect to see a

regime of diversification, followed by reduction as the new group disrupts the distribution. This shape is therefore a natural outcome of the process, and tracks with the conversation demonstrating that both economically and demographically the existing community has been displaced, leading to the collapse in the richness of the area. This outcome is most obvious in the demographic variables (figure 5). An inflection point in 2012 followed by four continuous years of decline, displaying this convex pattern. The index falls from .573 in 2012 to .557 in 2016 (with a local minimum of .551 in 2015). What is clear is that in both cases, both economically and demographically the areas surrounding The Beltline and The Eastside Trail are beginning to homogenize.

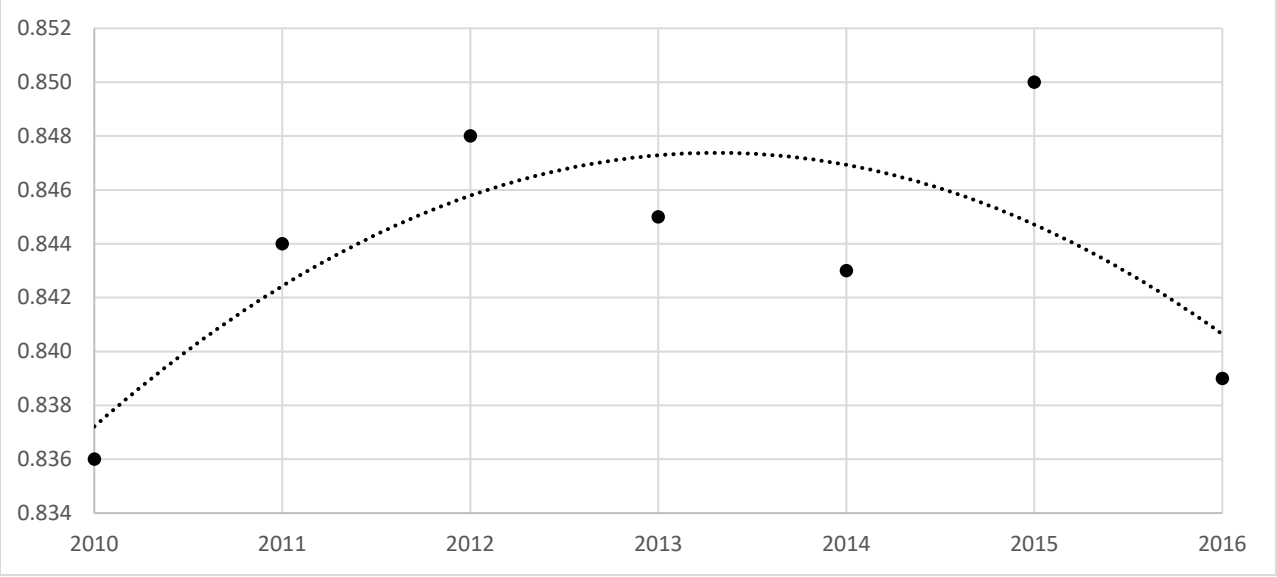

Figure 4: Graph showing average Shannon equitability (E) for economic variables for tracts in the study area from 2010 to 2016

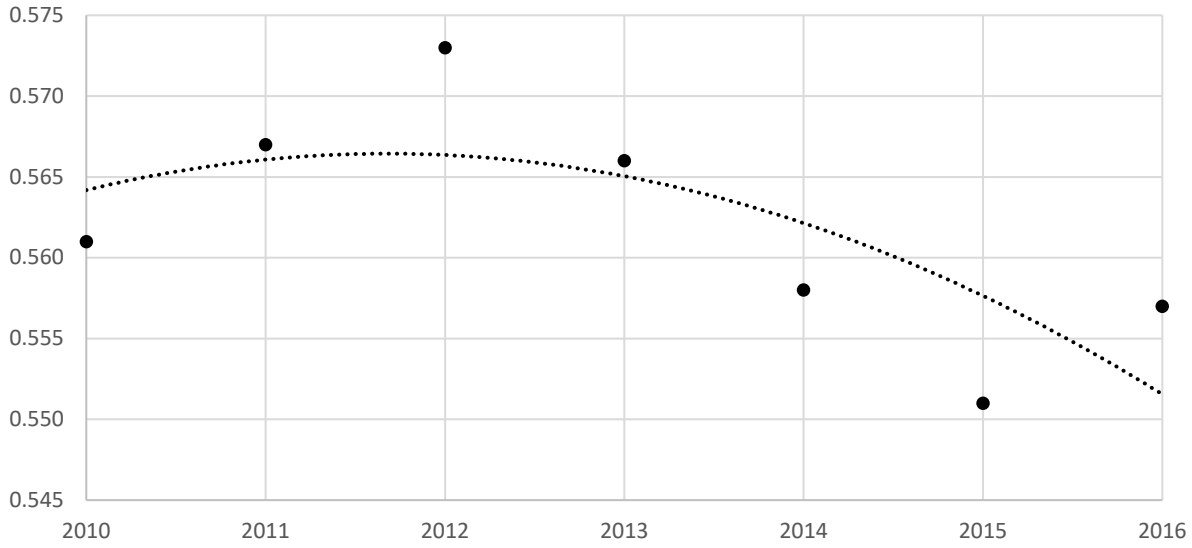

Figure 5: Graph showing average Shannon equitability (E) for demographic variables for tracts in the study area from 2010 to 2016

Discussion

While it is possible that these findings are specific to this area, or even specific to Atlanta, we are able to begin to extrapolate a generalization from these findings, as they do support the narrative of the zeitgeist, that The Beltline is seeing a reduction in diversity. There are two general statements that can be made, that moments surrounding an inflection point are critical times for intervention, and that the signs of the derivative of E with respect to some interval of time t is meaningful.

The sign of the derivative indicates the current regime. While the influence of magnitude requires additional research, the sign indicates a system gaining or losing in entropy. The slope then indicates how close the system is to regime change. If the slope is becoming less negative, if there is an inflection, this will change to a growth regime, where complexity and diversity increase. Conversely, if the slope moves from positive to negative across the inflection then the system is moving towards a regime of homogenization. More interestingly though is that this value can be used to determine this can be used as a planning tool. Examples of this are demonstrated visually in table 2.

This derivative value can be used to better deploy resources into an area. As the value approaches a critical point, as $\frac{dE}{dt} \to 0$, there is cause for attention to be paid. In either case, whether moving from a regime of growth to one of homogenization, a system displaying these traits is existing in a moment of instability. For those researching social phenomena, this is the moment to begin to document and research the forces of change that have led to this situation. For those responsible for policy, it provides focus and clarity on where issues may be emerging.

Using this interpretation, we are able to see that, for example, the tract containing Old Fourth Ward Park is approaching such a point with the rate of changing declining from .03 to .01 per year, when taken as a composite of both demographic and economic indexes. While the sign of the slope remains positive the decreasing rate of change portends the same parabolic change seen in the rest of the area. This is to suggest that currently, in terms of both demographics and economic social groups, diversity will likely begin to decrease. This is consistent with current trends in the area as new developments buy up properties previously inhabited by minority or marginalized groups, for example along Boulevard.

What has been demonstrated is that, by extracting metadata concerning the level of uncertainty and how those levels change over time, we are able to gain some insight into the processes of change in an area. The advantage over the extraction of descriptive statistics is that the function found in equation 5 is a relative measure, as opposed to an absolute one. If we wish to understand the phenomenology of gentrification, then we are less concerned with addition or removal in absolute terms, as we are with a purely economic description of gentrification. Rather, from a phenomenological lens, it is shifts in the relationships between factors that serve to alter perception and reshape the image of a place in the zeitgeist. The methodology as proposed and executed here serves to suggest a tool, and not a value that is in-and-of-itself wholly explanatory. The topic of gentrification, and the human experience of that phenomenon, is too complex to explain in such simple terms. Rather, by explaining the reductions or gains in complexity for variables that dominate the process, we describe, in relative and abstract terms, the level of social uncertainty a person experiences. Social uncertainty in this context is the rate at which, in the course of normal social life, one is confronted with someone whose lived experiences differ.

What is required though is a means to link these quantitative metrics to a qualitative understanding of the phenomenon. What we have provided is a means, much like uses for the SVI in identifying areas in need of disaster assistance[50], to narrow our focus to understand what areas are, for lack of a better term, at risk of this phenomenon occurring. While the index provides a means of identifying places that are at risk, it is itself not a key indicator. Without significant research and calibration of the model in a broader context, even just in the City of Atlanta, it is not possible to use this value as a means of declaring, absolutely, if a place is or is not gentrified or gentrifying. Rather this acts as a quantitative component to a mixed methods approach to investigating the phenomenology of gentrification, the perceptual and experiential factors that cause the image (perception) of a place to change.

There are of course limitations to the tool developed. While this measure has several interpretive benefits as discussed in earlier sections, due to the mathematics it also has limitations on how those results can be understood. This method provides metadata on multilevel categorical data on a variable by variable basis then aggregates across those variables in an unweighted manner. This unweighted average does not capture how different variables might impact perception in unequal ways. This could be rectified by the addition of a weight parameter that allows specific variables, as an example educational attainment, to count in an unequal manner towards the overall average. This would require significant study into the dynamics of the system and how subjective decisions are made. These variables also likely have significant interaction, creating a highly non-linear system. This creates problems in parameter estimation, making the process much more complex than the typical linear parameter estimation. This does though mean the index is read as though each factor in the analysis, per indicator, is equal.

It is tempting during this analysis to inquire as to how changes in a specific level of a variable impact the overall system. This method prevents this. Individual values per variable level (H_i) do not have meaning in isolation, or at least do not provide any additional information over traditional methods. H_i does not provide any additional information beyond what is provided by computing the percentage (p_i) required for the entropy calculation. Therefore, we cannot suppose

[50] Citation to paper concerning hurricane and climate change preparedness using SVI see CDC website for citation

from the output what effect changes in a specific variable level are having overall. The entropy value (H) is changing relative to the overall system. The value, for a specific variable, shows how much uncertainty there is in the outcomes of that variable. Since shifts in a given level are done at the expense of others, the parabolic nature of the equation means that the values do not add any additional information to the analysis. This is compounded by the fact that the growth of the equation is logarithmic leading to large initial gains, but relatively stable values as you approach 1/S. This weights first appearance and initial population growth more heavily than the addition of the n^{th} individual. The addition of the n^{th} individual must also be balanced out against a loss elsewhere in the population, thus any attempt at interpretation must take into account the changes in the overall distribution, which is what the overall value H represents.

Some of these limitations can be overcome by looking into other entropy calculations, other methods of representing the richness of the information content that may allow for the possibility of overcoming the directionless nature of H. As mentioned earlier, the value H represents the distance from an ideal uniform distribution across all levels of a variable. This leads to a lack of directionality in the index itself. While longitudinal analysis can show changes over time, and thus gains and losses, it still cannot show directionality at the case to case level. A measure that provided directionality would give each value Hi meaning in and of itself. It would indicate in which direction (over or under representation in the population) and relative amount (for any value pi greater than 0). This would again provide interpretive benefits. One possibility is to consider an application of Renyi entropy[51], as opposed to Shannon entropy.[52] Renyi entropy contains an order (or shape) variable, α, which can induce different behavior in the function. In the case of $\alpha = 2$ this becomes parabolic, giving the derivative of the function the ability to provide directionality.

Another representational shortcoming is the fact that the Shannon index does not discuss the concept in a spatialized way. In the current method a 'system' is bounded geographically by a

[51] Renyi 'On Measures of Information and Entropy' *Proceedings of the Fourth Berkley Symposium on Mathematics, Statistics and Probability* (1960) p.547-561.
[52] Shannon entropy is a special case of the Renyi equation where alpha is taken at unity ($\alpha = 1$). See A Renyi 'On Measures of Information and Entropy' *Proceedings of the Fourth Berkley Symposium on Mathematics, Statistics and Probability* (1960) p.550.

census tract, which does not truly encapsulate a system in its totality, since the boundaries are drawn with a consideration towards politics and demography. Providing some kind of spatial measure would be of interest in understanding and representing how, as one moves through a city, the potentials revealed by the index change across space, as much as across time by rearranging pi and instead using a ratio of densities over an even spatial increment. This would create a raster map showing what the level of social uncertainty is. This of course requires a finer grained understanding of the population and its distributions within a defined space.

This matter of spatiality is difficult for the current methodology, and acts as a major limitation. The variables selected are tabulated (for the most part) at the tract level. Tracts, even in urban areas, tend to be relatively large and this scale represents an issue, as noted above. Many variables are not available below this level due to confidentiality requirements. This level of resolution presents an issue due to an inherent assumption placed on the data, that all things are distributed evenly within the system (inside of a spatial extent). This is not necessarily true, and this assumption can, at the scale of the census tract, lead to some spurious assumptions or misinterpretations. For example, the census tract covering Old Fourth Ward covers from areas near The Beltline up to areas around Metropolitan Parkway, creating wildly different economic conditions. Inside of this coverage area there are multiple neighborhoods with different social systems and dynamics. This broad coverage could give the impression that there is heterogeneity in the area when, given local knowledge, there are multiple economically homogenous groups that do not generally mix. Ideally this would be overcome by use of data at the smallest possible census level (the block). The limitation arises in that not all variables are not available at this scale and are therefore inaccessible through normal means.

This problem of scale is not unique to this method. In 1993 Rubin brought this problem to light and proposed a solution in his article *Statistical Disclosure Limitation*.[53] Population synthesis represents a way to overcome the spatial scale limitation. The method uses survey data, such as the American Communities Survey (ACS) annual survey[54], or other microdata and projects that

[53] D Rubin "Discussion Statistical Disclosure Limitation" *Journal of Official Statistics* (1993) vol 9 p. 461-468
[54] This survey released as a part of the Public Use Microdata product. PUMs or PUMA data is an annual survey of 1% of the population of the United States of America and represents a high quality data product concerning the demographic makeup of the population. Information on the methods can be obtained at

data down to the desired scale. This is a common practice in data imputation and systems calibration where a known sample data set is used to stochastically simulate a unique data set. This data can be used to overcome the problem of scale. It is possible simulate a synthetic population for a small geography, such as a block, using PUMA data. Applications such as the Mobility Analytics Research Group PopGen2 package[55] or other packages can use this sample data to construct a synthetic population at an appropriate scale. Using data generated by this kind of simulation would allow the index a finer scale and allows us to dynamically address both coarse and fine grained phenomena. This would have the drawback of no longer being real data, and rather a representation of a representation, but these methods are well tested. The benefit of this approach would be the ability to create smallest possible scale, or in envelopes that better represent a neighborhood, community, or place[56].

What this method has generated, as mentioned, is an indicator that aids in the further development of a mixed quantitative and qualitative method to describe how gentrification is perceived to come about. This makes up a significant part of further work. While the statistical work of developing and deploying more robust versions of the index is important, the pairing of this methodology with a complimentary qualitative component that seeks to capture the human experience is most critical. While we have an intuition concerning where gentrification is happening, or where is 'next', this index is able to aid researchers in better allocating resources into the qualitative and ethnographic work that needs to be done. It serves, in a way similar to the SVI, to identify areas of the city that are beginning to experience changes in their social ecosystems, where we are beginning to see homogenization. There are, we believe, two critical research paths. The first, is further investigation of the patterns that are found in the data as revealed by the index. There is a paradoxical effect that occurs as gentrification or displacement begins, an observable increase in heterogeneity within the population. What happens around this critical point ($dH/dt = 0$) is important to understand and requires more investigation.

---

https://www.census.gov/geo/reference/puma.html and data may be obtained through the United States of America Census Bureau or through packages such as lodown for R.

[55] https://www.mobilityanalytics.org/popgen.html

[56] This assumes that the study boundary does not cross the arbitrary data tabulation boundary. In the case of this work it would require that any boundary be fully spatially contained (discretely within) a census PUMA.

This investigation though cannot be one that is based on a purely mathematical or quantitative understanding. Rather, there is a need for a mixed methodology that combines the quantitative approach with a more qualitative understanding of what happens as opinions and perceptions shift. The index is an abstraction of an experiential process, the way the image of an area changes over time, and that abstraction needs to be ‘fleshed out’. This requires complimentary work to explain the more subjective components.